Environment International 92–93 (2016) 617–629

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

# How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards

Andrew A. Rooney [a], Glinda S. Cooper [b], Gloria D. Jahnke [c], Juleen Lam [d], Rebecca L. Morgan [e], Abee L. Boyles [a], Jennifer M. Ratcliffe [f], Andrew D. Kraft [b], Holger J. Schünemann [e], Pamela Schwingl [f], Teneille D. Walker [b], Kristina A. Thayer [a], Ruth M. Lunn [c,*]

[a] Office of Health Assessment and Translation, Division of the National Toxicology Program, National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), Research Triangle Park, NC, USA
[b] National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC, USA
[c] Office of the Report on Carcinogens, Division of the National Toxicology Program, National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), Research Triangle Park, NC, USA
[d] University of California San Francisco, Program on Reproductive Health and the Environment, San Francisco, CA, USA
[e] McMaster University, Department of Clinical Epidemiology and Biostatistics, Hamilton, Ontario, Canada
[f] Integrated Laboratory Systems (ILS), Morrisville, NC, USA

article info

Article history:
Received 20 July 2015
Received in revised form 2 December 2015
Accepted 10 January 2016
Available online 6 February 2016

Keywords:
Risk of bias
Internal validity
Systematic review
Environmental health
Hazard assessment

abstract

Environmental health hazard assessments are routinely relied upon for public health decision-making. The evidence base used in these assessments is typically developed from a collection of diverse sources of information of varying quality. It is critical that literature-based evaluations consider the credibility of individual studies used to reach conclusions through consistent, transparent and accepted methods. Systematic review procedures address study credibility by assessing internal validity or "risk of bias" — the assessment of whether the design and conduct of a study compromised the credibility of the link between exposure/intervention and outcome. This paper describes the commonalities and differences in risk-of-bias methods developed or used by five groups that conduct or provide methodological input for performing environmental health hazard assessments: the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) Working Group, the Navigation Guide, the National Toxicology Program's (NTP) Office of Health Assessment and Translation (OHAT) and Office of the Report on Carcinogens (ORoC), and the Integrated Risk Information System of the U.S. Environmental Protection Agency (EPA-IRIS). Each of these groups have been developing and applying rigorous assessment methods for integrating across a heterogeneous collection of human and animal studies to inform conclusions on potential environmental health hazards. There is substantial consistency across the groups in the consideration of risk-of-bias issues or "domains" for assessing observational human studies. There is a similar overlap in terms of domains addressed for animal studies; however, the groups differ in the relative emphasis placed on different aspects of risk of bias. Future directions for the continued harmonization and improvement of these methods are also discussed.
Published by Elsevier Ltd.

## 1. Introduction

The assessment of study quality has long been considered an important part of synthesizing evidence to answer questions in toxicology and environmental health sciences (e.g., IARC, 1990; WHO, 1999). However, the term "study quality" is broad and can vary widely across the fields of systematic review and environmental health (e.g., see terminology discussion in Viswanathan et al., 2012). Recent initiatives in the environmental and occupational health community have emphasized the goal of increasing transparency and objectivity of the evaluation process by adopting systematic review methods (e.g., Birnbaum et al., 2013; EFSA, 2010; Woodruff and Sutton, 2014). As a result of these efforts, there is an increased focus on transparently evaluating one aspect of study quality — the assessment of systematic errors that can result in a biased (over- or under-estimated) effect estimate referred to as risk of bias or internal validity. Risk of bias is a measure of whether the design or conduct of a study alters the effect estimate or compromises the credibility of the reported association (or lack thereof) between

Abbreviations: AHRQ, Agency for Healthcare Research and Quality; NTP, National Toxicology Program; OHAT, Office of Health Assessment and Translation; ORoC, Office of the Report on Carcinogens; EPA-IRIS, Integrated Risk Information System of the U.S. Environmental Protection Agency; GRADE, Grading of Recommendations Assessment, Development, and Evaluation.

* Corresponding author at: NIEHS, P.O. Box 12233, Mail Drop K2-04, RTP, NC 2770; 530 Davis Drive, Morrisville, NC 27560, USA.
E-mail address: lunn@niehs.nih.gov (R.M. Lunn).

http://dx.doi.org/10.1016/j.envint.2016.01.005
0160-4120/Published by Elsevier Ltd.

exposure/treatment and outcome (Guyatt et al., 2011a; IOM, 2011; Viswanathan et al., 2012). The use of the risk-of-bias terminology has been supported by systematic review guidance groups such as the Cochrane Collaboration and the Agency for Healthcare Research and Quality (AHRQ) because it reduces ambiguity between the quality of reporting and the quality of the actual conduct of the research (Higgins and Green, 2011; Rooney et al., 2014; Viswanathan et al., 2012).

In this paper, we begin with a discussion of the application of systematic review methods to environmental health. We then present an overview of risk-of-bias approaches that have been developed or used to assess environmental health data by five different groups (the Grading of Recommendations Assessment, Development, and Evaluation [GRADE] Working Group; the Navigation Guide; the National Toxicology Program's [NTP] Office of Health Assessment and Translation [OHAT]; the NTP's Office of the Report on Carcinogens [ORoC]; and the Integrated Risk Information System of the U.S. Environmental Protection Agency [EPA-IRIS]). This analysis is based on discussions that occurred during 2014–2015 to address common interests in understanding, developing, or refining methods for assessing the credibility of individual studies as part of reaching conclusions on specific environmental health questions. Commonalities and differences in the approaches taken across the groups are highlighted along with a discussion of opportunities and challenges for harmonization as methods are refined and further developed over time. To ensure clear communication with a variety of scientific disciplines, definitions for terms commonly used in environmental health reviews and publications are provided (Table 1).

### 1.1. Application of systematic review methods to environmental health

A systematic review is a literature-based evaluation focused on a specific question that uses explicit, pre-specified methods to identify, select, assess, and synthesize scientific evidence (IOM, 2011). These methods increase the transparency, objectivity, and rigor in the review process. The systematic review methods being applied to environmental health questions have been built on the structure of established approaches for evaluating evidence in clinical medicine and public health, such as the Cochrane Collaboration (Higgins and Green, 2011), the Evidence-based Practice Center (EPC) methods guides for the AHRQ (AHRQ, 2013) and the GRADE Working Group (Atkins et al., 2004; Guyatt et al., 2011a). These approaches typically consider human evidence from different study designs (i.e., randomized controlled trials and observational studies) and have been applied widely to clinical medicine and public health.

Table 1
Definitions of common terms.

| Term | Definition |
|---|---|
| Domain (also used: Category or Question) | Issue or topic within risk of bias such as "confounding" or "selective outcome reporting" |
| Indirectness (also used: Applicability or external validity) | Measure of how well a study addresses the specific question of the systematic review or the extent to which results inform the review question |
| Reporting quality (also used: study quality) | Measure of how thoroughly details on study design, experimental procedures, results and analyses were reported (Reporting only addresses a portion of the larger concept of Study Quality; however, sometimes the terms are conflated) |
| Risk of bias (also used: internal validity, study quality) | Measure of the credibility of study findings that reflects the ability of a study's design and conduct to protect against systematic errors that may bias (over- or under estimate) the results or estimate of effect (Risk of Bias only addresses a portion of the larger concept of Study Quality; however, sometimes the terms are conflated) |
| Sensitivity | The ability of a study to detect a true risk (similar to the concept of a sensitive assay); an insensitive study will fail to show a difference that truly exists, leading to a false conclusion of no effect. Example considerations include having adequate numbers of exposed cases, exposure levels, durations, ranges, windows of exposure, and lengths of follow-up. |
| Study quality | A complex idea with different meanings for different groups including one or more of the following: reporting quality, applicability and risk of bias. For systematic review methods study quality generally includes risk of bias assessment. |
| Systematic review | A review of literature focused on a specific question that uses explicit, pre-specified methods to identify, select, assess, and synthesize scientific evidence |

There is considerable variability in the study designs and data sources available to evaluate potential health effects from exposure to environmental chemicals, necessitating some modification of methods developed in clinical medicine. Unlike questions in clinical medicine, environmental datasets rarely include controlled human exposure studies because ethical considerations generally rule out exposing human subjects to chemicals suspected to pose a health hazard. When available, controlled human exposure studies are typically limited to short-term exposures and temporary or reversible health endpoints such as the series of investigations on inflammatory and cardiovascular indicators associated with exposure to diesel exhaust (see Ghio et al., 2012); these types of studies may be of limited relevance to questions regarding effects of longer term exposures. Studies of "natural experiments" wherein researchers take advantage of unplanned exposures or external factors that interrupt exposure [e.g., reduced air pollution associated with the Beijing Olympics allowing an examination of the impact of air pollution on birth weight (Rich et al., 2015)], can provide another useful source of human health effects data (Craig et al., 2012). However, availability of such data is very limited. More typically, human data are derived from a variety of observational designs, including cohort studies, case–control studies, and clinic-based or population-based surveys, as well as from ecological studies or case series or reports.

Questions in environmental health often require the assessment of a broad range of relevant data including animal and mechanistic studies as well as human studies. Experimental animal data, primarily from in vivo laboratory studies in rodents, provide a large proportion of the toxicology data used for hazard identification and risk assessment. Studies of wildlife or animals living in heavily contaminated sites using an observational design may provide health effect data for chemicals that are widely distributed in the environment. Mechanistic data can be found in a wide variety of in vitro and in vivo studies, or studies of molecular, biochemical and cellular events in humans, rather than studies of the disease phenotype (i.e., molecular epidemiology studies). These data may explain how a chemical produces particular adverse health effects and can inform the hazard conclusions.

For environmental health questions, the most widely available in vivo data generally come from experimental animal and observational human epidemiology studies. Whatever the evidence base is, critical assessment of individual studies is needed to evaluate each of the evidence streams (human, animal, and mechanistic studies) with clear consideration of the strengths and weaknesses of different study designs.

## 2. Overview of current methods (frameworks and tools)

The five groups are involved in conducting systematic reviews that may differ in focus (e.g., cancer or non-cancer endpoints; short term or lifetime hazard evaluations; derivation of risk estimates), scope (individual health endpoints or comprehensive toxicological evaluations; simple or complex literature databases considered), underlying guidance (e.g., agency guidelines that must be adhered to), and use of the systematic reviews by regulatory agencies. The approach taken for evaluating risk of bias and incorporating that evaluation into the systematic review should match the intended purpose of the review for the organization involved. For example, the product of an OHAT systematic review will vary depending on the question and the extent of the available

evidence, and may take the form of NTP hazard identification conclusions, opinions on whether substances may be of concern given what is known about toxicity and current human exposure levels, or state-of-the-science reports that do not include formal NTP conclusions. The application of EPA systematic reviews can range from complex IRIS assessments including hazard identification and dose–response analyses that can be used as a basis for setting long-term regulatory standards to much more focused reviews needed in a very short timeframe to temporarily inform an environmental cleanup.

In general, the approaches for assessing risk of bias are similar across the five groups (see Table 2 for details on the current methods for each framework). Ratings are developed for individual studies on separate risk-of-bias issues or "domains" that may compromise the credibility of the reported association (or lack thereof) between exposure/treatment and outcome based on criteria regarding the study design, conduct, and reporting. The ratings for each domain reflect a judgment of the potential bias for that domain using a scale to categorize the extent of bias (e.g., high, medium, or low). Each group uses a framework in which the

Table 2
Overview of five frameworks for evaluation of risk of bias of environmental health studies.

| Group | Scope | Approach | Experience to date |
|---|---|---|---|
| GRADE | Assessment of internal validity for randomized and nonrandomized studies of interventions | General approach:<br>• Assessment occurs on an outcome basis for each study and then across studies for a specific question | GRADE RoB criteria have been applied to studies of environmental health trials and interventions. More research is needed on the application to studies of environmental exposures |
| Navigation Guide | Assessment of chemicals with the goal of expediting the development of evidence-based recommendations for preventing harmful environmental exposures | General approach:<br>• Separate methods for application to specific study designs<br>• Core question for each domain, accompanied by description and examples for each possible rating<br>• Consider direction of bias/limitation if possible<br>Human observational studies (applies to wildlife/animal observational studies):<br>• 9 domains (exposure, outcome, selection, confounding, blinding, incomplete outcome data, selective outcome reporting, financial conflict of interest, other)<br>Animal toxicology (experimental) studies:<br>• 7 domains (sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting, financial conflict of interest, other) | Two case studies published in peer-reviewed journal; one case study in preparation to be submitted to peer-review journal; two case studies in progress with protocols published in PROSPERO |
| OHAT | Assessment of the evidence that environmental chemicals, physical substances, or mixtures cause adverse non-cancer health effects and provides opinions on whether these substances may be of concern given what is known about current human exposure levels | General approach:<br>• Single set of questions with subsets applied to specific study designs<br>• Separate criteria, description and examples for each rating by study design<br>• Assessment occurs on an outcome basis for each study<br>• Consider direction of bias/limitation if possible<br>Human observational studies (applies to wildlife/animal observational studies):<br>• 7 domains [exposure, outcome (includes blinding of outcome assessors), selection, confounding, attrition/exclusion, selective outcome reporting, other]<br>Animal toxicology (experimental) studies:<br>• 9 domains [randomization, allocation concealment, identical experimental conditions, blinding during study, exposure, outcome (includes blinding of outcome assessors), attrition/exclusion, selective outcome reporting, other]<br>Also tailored approaches for human controlled trials and in vitro exposure studies | Multiple assessments in process. All evaluations follow similar process (http://ntp.niehs.nih.gov/go/38,138) with opportunities for external scientific, interagency, and public input. Protocols posted on NTP webpages (http://ntp.niehs.nih.gov/go/evals). |
| RoC | Assessment of carcinogenicity of chemicals for listing in the Report on Carcinogens (RoC). The RoC is a congressionally mandated, science-based, public health report that identifies agents, substances, mixtures, or exposures in our environment that pose a cancer hazard for people in the United States | General approach:<br>• Separate methods developed by discipline<br>• Considers direction and impact of bias/limitation if possible<br>• Overall evaluation of the ability of each study to inform the hazard evaluation<br>Human epidemiology studies:<br>• 7 domains: 6 for risk of bias (selection and attrition bias, information bias from exposure misclassification, information bias from outcome misclassification, potential for confounding, analysis, selective reporting) and 1 for sensitivity<br>• Series of signaling and following questions used to provide a rating for a core question for each domain<br>Animal toxicology studies:<br>• 5 categories (study design, exposure conditions, outcome assessment and measurement, potential for confounding, analysis and selective reporting). Each domain consists of risk of bias and sensitivity signaling and follow-up questions | Similar methods applied to evaluation of human and/or animal studies for preparation of RoC monographs for four substances, which were reviewed by an external peer review panel in a public forum with opportunity for comment. Draft RoC monograph for substances in progress (http://ntp.niehs.nih.gov/go/37894). |
| EPA-IRIS | Assessment of toxicity of chemicals and other environmental exposures; includes cancer and non- cancer (e.g., reproductive, developmental, neurotoxicity, immunotoxicity) evaluations of hazard and provides quantitative dose–response estimates. Used in EPA decisions | General approach:<br>• Separate methods developed by discipline<br>• Consider direction of bias/limitation if possible<br>Epidemiology studies:<br>• 7 domains of biases or limitations (exposure, outcome, selection, confounding, analysis, selective reporting, sensitivity)<br>• Protocol-based evaluation of each domain; prompting questions used to guide review<br>Animal toxicology studies:<br>• 4 study features (experimental design, exposure, endpoint, outcome reporting and analysis)<br>• Each evaluated for bias and sensitivity<br>• Series of signaling and follow-up questions for each domain | Application to multiple assessments in process; review by external peer review panel (with additional opportunity for public review and comment) |

risk-of-bias tool is tailored for specific study designs (e.g., randomized clinical trials versus observational studies versus experimental animal study). Risk of bias is assessed on an outcome basis because different outcomes may have been measured with methods that differed in their accuracy, objectivity, reliability, or sensitivity. Reviewers are encouraged to identify the direction of bias when possible.

The groups all develop project-specific risk-of-bias criteria in a protocol to guide development of risk-of-bias ratings for each question or domain prior to conducting the systematic review and use topic-specific experts to provide input on drafting or reviewing the risk-of-bias criteria (e.g., expertise in the exposure or outcome assessment methods under review). The risk-of-bias criteria describe aspects of study design, conduct, and reporting required to reach risk-of-bias ratings for each domain (e.g., what separates low bias from medium bias). Each of the groups also recommends that a small subset of studies be included in a "pilot" phase to discuss and resolve any ambiguity before proceeding with evaluation of the full set of studies. For the full evaluation, the groups use a minimum of two independent reviewers and then determine the final risk-of-bias rating through discussion and consensus, or third-party consultation where there are disagreements.

There are a number of different ways the risk-of-bias ratings from the individual studies can be used in later steps of a systematic review, and there are differences among the five groups in how risk-of-bias assessments are incorporated in their frameworks (see Fig. 1). Applications, as discussed below, include: to interpret the findings from individual studies; to identify the most informative studies or to exclude studies with the highest level of bias; or as a factor used to evaluate certainty (also referred to as confidence or strength) of the evidence across the body of studies.

The risk-of-bias assessment can be applied in the interpretation of the findings of individual studies. That is, the confidence in a study's findings depends on a rigorous risk of bias evaluation for all domains in conjunction with the strength of the observed association between exposure to the substance and the health outcome for each study. The presence of potential bias in a study does not necessarily mean that the study should be excluded from the assessment. For example, the level of concern for the potential for bias (or probability of bias), the range of different types of biases and also the direction and degree of distortion of the effect estimate from potential biases (in the different domains) can impact how studies are considered for hazard identification. The interpretation of results from a study with potential biases toward the null (such as non-differential exposure assessment or healthy worker effect) would differ from the interpretation of results from a study with potential biases away from the null (such as potential for confounding or recall bias). Low or weak risk estimates in the former case may provide support for an association between the exposure and outcome of interest, whereas even positive findings in the latter case may be considered to be inconclusive. Similarly, the magnitude of bias can have a major effect on how individual studies are considered in reaching conclusions. The results of studies with the highest level of bias may be entirely due to bias; whereas, more rigorous studies are more likely to produce findings that are closer to the truth. In some cases it may be possible to calculate or estimate the degree of confounding or distortion of a bias on the effect estimate, by performing an analysis that removes studies with particular biases to investigate changes in the overall effect size. For example, it may be possible to conclude that the potential for confounding or bias would only explain some but not all the excess risk reported in a study depending on the magnitude of the effect estimate and of distortion of the bias.

Risk of bias can also be used to identify studies that may be given greater weight in reaching conclusions on health hazards. There are several ways risk-of-bias ratings can be used to identify these studies. One approach is to exclude studies with the highest level of bias (e.g., a "critical" risk of bias as described by Sterne et al., 2014). Another approach is to sort studies by an overall study-level rating on the ability of the study to inform the evaluation or as a way to stratify the analysis to see if the results are similar across different groupings of studies (note that these two approaches are not mutually exclusive). These ratings are based on the assessment of potential bias for all of the domains examined, although for some groups this rating reflects risk of bias (Navigation Guide, OHAT, GRADE) and for others each domain is evaluated for bias and study sensitivity (ORoC, EPA-IRIS; see Table 2). The overall judgment of risk of bias, or for some groups a larger concept of "study utility" (ORoC, EPA-IRIS) is not meant to be an algorithm that sums up the ratings across domains; ratings for the different domains may be given greater emphasis depending on the scientific issues important for the evaluation of the specific substance under review.

A third application of the risk-of-bias assessment is as one of several factors used to evaluate certainty (also referred to as confidence or strength) in the evidence across a body of studies. When the evaluation includes a meta-analysis, risk of bias across studies is also informative in evaluating the confidence in a summary effect estimate, similar to interpreting the findings from individual studies based on their strengths and limitations. Risk of bias across studies by outcome is one criterion that may be used to rate down certainty in the body of evidence (and then overall certainty across outcomes) in the GRADE approach (see the third column of Fig. 1). The Navigation Guide and OHAT systematic review methods also use the GRADE approach, with some modifications. ORoC and EPA-IRIS use a set of considerations that overlap with those used in the GRADE framework, and begin with the synthesis of the results from the higher confidence studies.

### 2.1. GRADE

Risk of bias is one of eight factors used within the GRADE approach to assess the overall certainty in a body of evidence across outcomes (see the third column of Fig. 1) (Atkins et al., 2004; Balshem et al., 2011). Rather than develop a specific risk-of-bias tool for the assessment of individual studies, GRADE currently suggests that risk of bias should be assessed using tools appropriate to the design of the studies included in the body of evidence. The GRADE approach highlights limitations to consider for randomized and non-randomized studies when assessing the risk of bias, as threats to risk of bias can reduce the overall certainty across a body of evidence (Guyatt et al., 2011c). Risk of bias is assessed for each study using Cochrane or other risk-of-bias tools and across studies for each outcome because overall risk-of-bias rating can differ across the outcomes. The across study risk-of-bias assessment of the body of evidence can either identify "no serious limitations", "serious limitations", or "very serious" limitations. A judgment is then made whether the bias identified is serious enough to rate the certainty of the evidence down one or two levels. The implication of these three levels are that for no serious limitations most information is from studies at low risk of bias, for serious limitations most information is from studies at moderate risk of bias, and for very serious limitations most information is from studies at high risk of bias; however, this interpretation may be modified with the recent release of A Cochrane Risk of Bias Tool for Non-randomized Studies of Interventions (ACROBAT-NRSI) (Schünemann et al., 2012; Sterne et al., 2014).

### 2.2. Navigation Guide

The Navigation Guide uses the GRADE framework for assessing certainty, or strength, of a body of evidence. At the individual study level, the Navigation Guide approach consists of separate but parallel risk-of-bias tools for evaluating human and animal evidence. Risk of bias is assessed using an adapted instrument based on existing guidance used by the Cochrane Collaboration (Higgins and Green, 2011) and the Agency for Healthcare Research and Quality (Viswanathan et al., 2012) for evaluating risk of bias of evidence in the clinical sciences. The Navigation Guide also considers the funding source of included

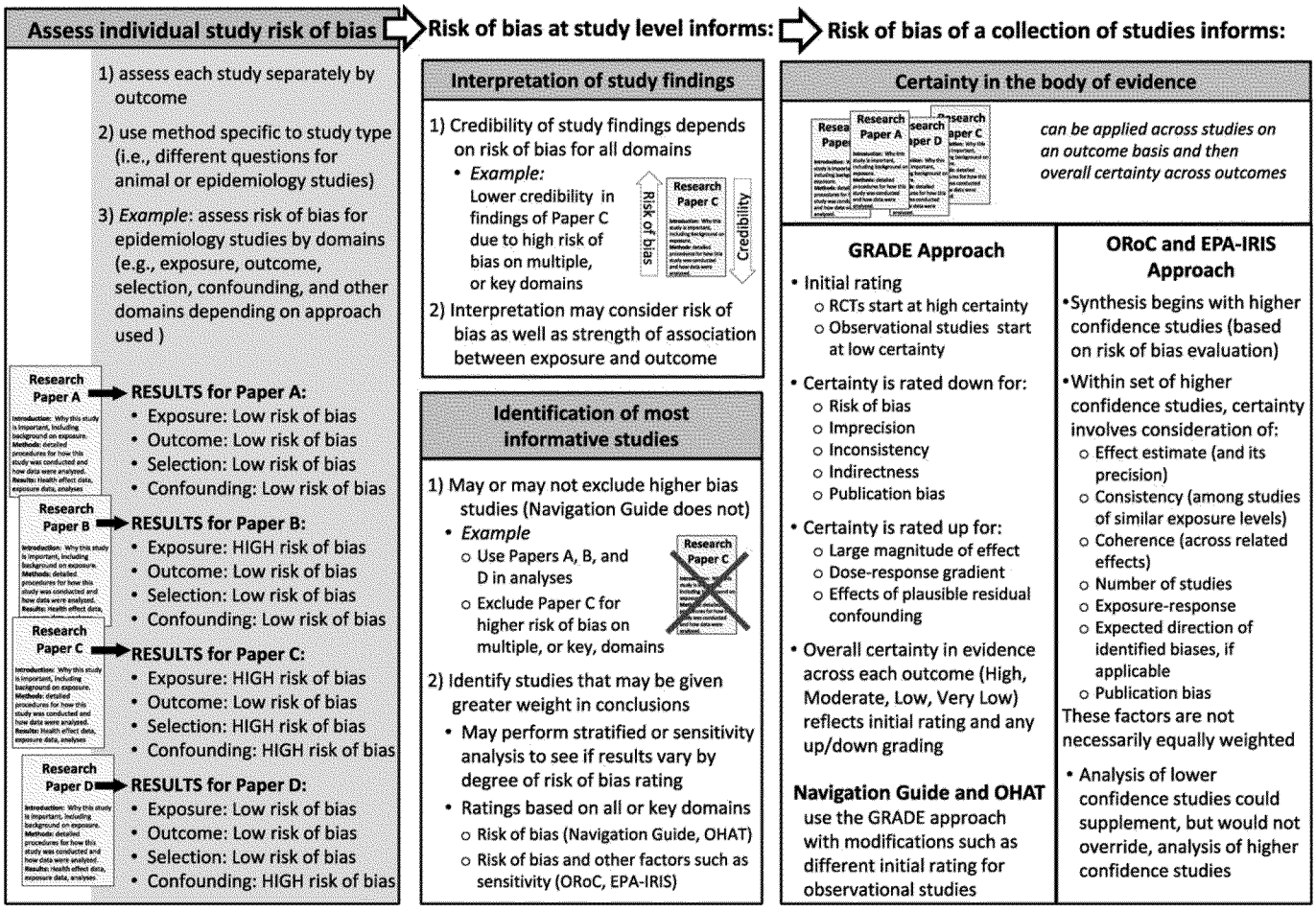


Fig. 1. Risk of bias of individual studies and its use in the evaluation of the body of evidence.

studies and evaluates financial conflicts of interests as a risk of bias, based on empirical data from studies conducted on pharmacological treatments that report evidence of bias associated with funding source (Krauth et al., 2013; Lundh et al., 2012). The risk-of-bias tool consists of overarching questions for each type of bias, each followed by detailed instructions outlining risk-of-bias criteria or considerations to incorporate when determining the rating. Although the overarching questions for each risk-of-bias domain applies broadly and are intended to be used across different systematic reviews, separate risk-of-bias criteria and instructions may be developed for each review to address anticipated issues specific to the study question at hand. When evaluating risk of bias at the individual study level, each domain is rated with one of five possible options: "low," "probably low," "probably high," "high," and "not applicable." These individual-study ratings are then considered across studies, as one component for assessing the overall certainty in the body of evidence in the GRADE approach. Modifications have been made to the GRADE approach for the initial certainty of observational studies to start at Moderate.

### 2.3. OHAT

OHAT also uses the GRADE framework for assessing certainty, or confidence, in a body of evidence. The current OHAT risk-of-bias tool for individual studies takes a parallel approach to evaluating risk of bias from human and animal studies to facilitate consideration of risk of bias across evidence streams with common domains and terminology (NTP, 2015b, c). The OHAT risk-of-bias tools are consistent with methods used by the Navigation Guide and other groups or recent guidance recommendations (Bal-Price and Coecke, 2011; Higgins and Green, 2011; Krauth et al., 2014; Liberati et al., 2009; McPartland et al., 2014; NTP, 2013a, b; Viswanathan et al., 2012). The current OHAT approach is not to consider conflict of interest as a domain of risk of bias; this factor is considered as part of evaluating publication bias across a body of studies. Individual risk-of-bias questions are designated as only applicable to certain study designs (e.g., cohort studies or experimental animal studies), and a subset of the questions apply to each study design. Criteria and instructions describing aspects of study design and conduct that determine risk of bias ratings are tailored for evidence stream and study design. When assessing internal validity of individual studies for a given outcome, each domain is rated with one of four options: "definitely low," "probably low," "probably high," or "definitely high" risk of bias. The risk-of-bias ratings of the entire collection of studies on a health outcome are then considered as one factors in assessing the strengths and weaknesses of the evidence for developing confidence ratings in the body of evidence. OHAT uses the GRADE framework for rating confidence (Guyatt et al., 2011a; Rooney et al., 2014) with modifications on initial starting point for observational studies and consideration of consistency across species, study designs, or human populations as an additional factor that may increase confidence in the association of exposure and health outcome.

### 2.4. ORoC

The ORoC's process for evaluating human epidemiology studies and animal cancer studies uses a series of questions tailored to the substance under review related to internal validity (e.g., potential for bias) and

study sensitivity (i.e., the ability of a study to detect a true risk or hazard, see Cooper et al., 2016–in this issue); animal studies are also evaluated for external validity (i.e., applicability of the model or results to the review question) (NTP, 2015a). For human studies, the questions are grouped into domains (for a specific type of bias or inadequate sensitivity), and the potential for bias in each domain is captured by a core question that expresses the underlying concerns regarding each type of bias. In general core questions are the same across designs although some signaling and follow-up questions (such as those dealing with selection bias) may vary by design. For animal studies, responses are made for each question in a specific domain (e.g., study design or exposure conditions) and questions related to sensitivity may be considered in several categories. The responses to relevant questions include "low/minimal concerns," "some concerns," "major concerns," "critical concerns," or "no information" and are based on guidelines developed from background research on the specific substance and issues (such as specific type of exposure assessment) related to the substance. When there is adequate information, a judgment is made on the direction of the potential bias (over- or under-estimate of the effect estimate, or unknown) and the potential magnitude of the distortion of the bias on the effect estimate. The concept of study utility is used to identify the most informative studies and interpret the findings from the studies. The overall evaluation of study utility is based on integration of the assessments for the domain-level judgments and the most informative studies are given greater weight in the conclusions. The identification of the potential for specific types (e.g., each domain) of uncontrolled bias or confounding and the assessment of study sensitivity are also used to rate confidence in the findings from studies and to help explain heterogeneity across studies. The evidence is synthesized across studies using the RoC listing criteria to determine the level of evidence conclusions from cancer studies in humans and animals. Several of the Hill factors (Hill, 1965) are considered in reaching level of evidence conclusions; however, it should be noted that these factors are not required in order to demonstrate causality (Rothman and Greenland, 2005).

### 2.5. EPA-IRIS

EPA-IRIS is developing a process for evaluating epidemiology and animal toxicology studies using specified classification criteria based on considerations of specific aspects of a study's design and conduct. To the extent possible, the evaluation will take into account the severity and anticipated impact of noted deficiencies. The criteria are developed based on background research pertaining to specific issues concerning the studies under review. In addition to the rating of specific domains (e.g., exposure measures), each study (or a specific analysis in a study) would be classified as "high," "medium," "low," or "not informative" with respect to confidence in the results. The principles and framework used for the evaluation of epidemiology studies examining chemical exposures are based on the recently developed ACROBAT-NRSI, Sterne et al., 2014), modified to address the exposure assessment and analysis issues typically encountered in occupational and environmental epidemiology research. The evaluation of the animal toxicology studies, rather than being organized as an assessment of risk-of-bias categories, focuses on an assessment of each component of the experiment (experimental design, including choice of animal model; exposure methods; endpoint evaluation methods; and outcome reporting and analysis). This approach was chosen to provide a organizational structure that fully covers the issues arising in toxicology research. These experimental features are then evaluated for bias and sensitivity. This approach is an adaptation of other published methods, and draws upon the breadth of issues of interest (specifically, issues relating to study sensitivity, see Cooper et al., 2016–in this issue) included in the Science in Risk Assessment and Policy (SciRAP) evaluation process, which relies extensively on reporting quality to determine study reliability (Agerstrand et al., 2011; Beronius et al., 2014; Mollenhauer et al., 2011). Similar to the approach described above for ORoC, the results of the individual study evaluations are used to identify the most informative studies, and are considered in the synthesis of the body of evidence as depicted in Fig. 1.

### 2.6. Commonalties and differences across methods

#### 2.6.1. Similarities among the frameworks.

There is substantial consistency in the consideration of risk-of-bias domains for observational human studies across the five groups (see Table 3 for brief descriptions of the domain coverage for each group). All of the groups evaluate participant selection, confounding, attrition/exclusion, exposure/intervention assessment, outcome assessment and selective reporting. Although the same risk-of-bias issues are covered, some of the domains are defined slightly differently across the groups (e.g., selection and attrition/exclusion), and there is some variation in what elements are included in which domains. For example, several groups assess attrition/exclusion as part of the selection bias domain (Navigation Guide, EPA-IRIS, ORoC) whereas OHAT evaluates it in a different risk-of-bias domain. The number of domains and the placement of risk-of-bias elements within a particular domain are unlikely to influence the individual study ratings because none of the groups develop risk of bias "scores" reflecting a sum or average rating across domains. One of the reasons that summary scores are discouraged in assessing risk of bias is that a score would be influenced by the number of elements and would not account for potential differences in the relative importance across domains (Higgins and Green, 2011).

Most of the groups put considerable emphasis on the evaluation of exposure measures. Confidence in the exposure characterization requires valid, reliable, specific, and sensitive methods that are applied consistently and that can distinguish between exposed and non-exposed people or among exposure categories at a relevant window of exposure. The confidence in the exposure characterization typically involves an evaluation of the quality (e.g., reliability and validity) of the exposure assessment methods and information on the exposure setting (e.g., workplace conditions with high exposure to all individuals). Quantitative estimates of each individual's exposure to the substance of interest based on relevant or multiple metrics are ideal, but qualitative measures which allow for the separation of exposure categories to draw inferences regarding relative risk can also be acceptable. Exposure misclassification or measurement error may be independent of the outcomes (non-differential) or related to the outcome of interest (differential). Non-differential measurement error of exposures will usually bias the results toward the null by lowering precision and therefore reducing the ability to distinguish potential effects between non-exposed and exposed subjects or among different exposure categories. Differential measurement of exposures across the exposure groups will also bias the exposure-outcome relationship, although the direction of the biases is less clear; some examples include observer and recall bias (Blair et al., 2007; Christensen et al., 2014).

The risk-of-bias domains for experimental animal studies used by the 5 groups are summarized in Table 4. As with the human studies, there is considerable overlap in terms of what is covered. While addressing similar elements, the groups differ in terms of where elements are placed across domains, emphasis on different issues considered under risk of bias, and the overall organization. All of the groups evaluate study design, blinding, attrition/exclusion, outcome assessment and selective reporting. The study design and conduct features considered under the study design domain reflect the broadest range of considerations across groups and includes issues such as randomization and allocation concealment that are sometimes treated as separate domains. For GRADE the study design domain focuses on all criteria considered based on the study design (i.e., for randomized trials, randomization, allocation concealment and blinding of those applying the exposure are critical factors within this domain). For the Navigation Guide, the study design domain focuses primarily on randomization and allocation concealment for experimental animal studies similar to the consideration of these study design issues for risk-of-bias tools that address

RCTs, or "experimental" studies in humans. OHAT, ORoC and EPA-IRIS include issues beyond randomization such as blinding of outcome assessment, housing practices, consistent use of treatment vehicle, and selection of study animals. Exposure characterization is also considered across most methods. For experimental studies, the ability to control exposure generally minimizes differential errors. However, accuracy of the exposure characterization, including purity and stability of the test agent for controlled exposure studies and potential for background contamination from caging or diet, is important to reduce non-differential measurement errors, particularly for low-dose studies, or where there may be concerns for impurities in the dose preparation.

Table 3
Consideration of domains across five methods for assessing risk-of-bias of human observational studies.

| Domain | GRADE[a] | OHAT[b] | Navigation Guide | RoC | EPA-IRIS |
|---|---|---|---|---|---|
| Participant Selection | • Assesses whether eligibility criteria developed and applied appropriately (e.g., under- or overmatching in case–-control studies, selection of exposed and unexposed in cohort studies from different populations) | • Assesses whether selection of participants results in appropriate comparison groups (e.g., from same population and using same eligibility criteria; cases and controls similar other than disease status) | • Evaluates eligibility criteria, recruitment and enrollment procedures, participation and follow-up rates across exposure or outcome groups | • Assesses selection into (or out of) study that is related to both exposure and outcome<br>• Considers healthy worker effects (hire or survival) or other types of healthy participants | • Considers selection into (or out of) study or analysis jointly related to exposure and to outcome |
| Confounding | • Considers adequate control of confounding, measurement of all known prognostic factors, matching for prognostic factors, adjustments in statistical analysis | • Assesses adequacy of adjustments or explicit considerations made for primary covariates and confounders in the final analyses (e.g., matching, statistical adjustment) | • Evaluates whether study appropriately assessed and accounted for all important confounders (lists of important confounders are developed beforehand with topic experts) | • Adequacy of the method or other information to address confounding considered<br>• Overall assessment of confounding considered in discussion of findings | • Confounding adequately addressed by various methods (matching, statistical adjustment, lack of associations, other)<br>• Consideration of over-adjustment |
| Attrition/exclusion | • Assesses if loss to follow-up adequately addressed | • Considers if loss of subjects adequately addressed | • Considered as part of selection bias domain | • Considered as part of selection bias domain | • Considered as part of selection bias domain |
| Exposure/intervention assessment | • Assesses measurement of the exposure, including differences in the measurement of exposure (e.g., recall bias in case–-control studies) | • Evaluates consistency of exposure assessment (e.g., same method, time frame)<br>• If relevant time window for outcome<br>• If range and variation sufficient to distinguish levels of exposure<br>• Considers use of values relative to limit of detection | • Considers exposure assessment accuracy: presence of exposure misclassification and appropriate QA/QC<br>• Separate question to address blinding: whether exposure assessors were blinded to outcome | • Considers ability to distinguish between (i.e., classify) exposed and non-exposed people, exposure or different exposure categories<br>• Considers relevant window and metric of exposure<br>• Considers observation and recall bias | • Considers if relevant time window of exposure for outcome<br>• Assesses ability to distinguish levels of exposure<br>• Considers reverse causality<br>• Considers use of values b limit of detection |
| Selective reporting | • Considered as part risk of bias assessment by outcome | • Assesses whether results were provided for all relevant pre-specified measures and for all participants | • Assesses whether outcome data for all participants is reported or appropriate statistical methods are used to impute missing data<br>• Assesses whether selective outcome reporting is an issue (i.e., all pre-specified outcomes reported) | • Considers if results are provided for all relevant measures and participants | • Assesses if results provided for all relevant measures and participants |
| Outcome assessment | • Evaluates measurement of the outcome, including differential surveillance for outcome in exposed and unexposed in cohort studies | • Evaluates whether outcome was assessed with valid and reliable method applied consistently (e.g., same method and length of time)<br>• Also assesses whether outcome assessors were blinded to treatment | • Assesses whether outcomes were assessed and defined consistently across all study participants using reliable methods with appropriate sensitivity analyses<br>• Separate question to assess blinding: whether outcome assessors were blinded to exposure | • Considers the ability to distinguish between the presence or absence (or degree of severity) of the outcome<br>• Considers whether misclassification varied across exposure group<br>• Considers observation bias | • Considers blinding<br>• Considers sensitivity and specificity of disease (outcome) measures; ability to distinguish presence or absence (or degree of severity) of disease (outcome) |
| Conflict of interest | • Issues related to conflict of interest assessed on many levels: publication bias, selective outcome reporting bias, and through conflict of interest management at the evidence to decision level | • Addresses conflict of interest elsewhere in methods, outside risk of bias | • Financial conflict of interest addressed | • Not addressed | • Not addressed |

(continued on next page)

Table 3 (continued)

| Domain | GRADE[a] | OHAT[b] | Navigation Guide | RoC | EPA-IRIS |
|---|---|---|---|---|---|
| Analysis | • Addressed under risk of bias when study level data is used but ideally dealt with through re-analysis of original data | • Analysis and statistical methods addressed under other potential threats to internal validity | • Analysis and statistical methods addressed under domain for Other | • Considers data assumptions and analysis adequate or that the study did not conduct relevant analysis of the available data | • Considers analysis strategy and details |
| Sensitivity | • Not addressed | • Exposure and outcome sensitivity addressed under questions for Exposure and Outcome assessment domains | • Exposure and outcome sensitivity addressed in Exposure and Outcome assessment domains | • Considers factors that could affect the ability to detect a true risk such as the number of exposed cases, exposure level duration, and range, and length of follow-up | • Considers sensitivity of design or methods (other attributes that could affect ability to detect true risk) |
| Other | • Considers other limitations such as early stopping for benefit<br>• Approach is based on, and influence by, the question that is asked | • Questions added on a project-specific basis to address potential threats to internal validity not addressed elsewhere (e.g., inappropriate statistical methods) | • Considers:<br>◦ Early stopping due to data-dependent process<br>◦ Claim of fraudulence<br>◦ Selective reporting of subgroups | • Not addressed | • Not addressed |

[a] Assessment occurs on an outcome basis for each study and then, across studies for a specific question; See Guyatt et al. (2011c) for details on the GRADE approach to assessing internal validity.

[b] See http://ntp.niehs.nih.gov/go/38673 for the current OHAT risk-of-bias tool.

#### 2.6.2. Differences among the frameworks

As noted above, some of the differences in the frameworks are relatively minor, and generally result from the placement of elements in different categories or domains. There are also differences in how the risk-of-bias assessment is used as part of an evaluation. For example several groups exclude studies with the highest level of bias (e.g., a "critical" risk of bias as described by Sterne et al., 2014 is used in the ORoC and EPA-IRIS approaches; or the high risk of bias "tier" of studies in the OHAT approach). GRADE suggests conducting sensitivity analyses (i.e., comparing results of low versus high risk of bias studies), ideally based on a priori defined criteria. Other approaches, such as the Navigation Guide, do not exclude studies based on risk-of-bias issues but would incorporate these findings during evaluation of the overall body of evidence, for example by downgrading the contribution of these studies in the overall strength or quality ratings. The OHAT and Navigation Guide methods for animal studies are organized around risk of bias; whereas, the EPA-IRIS approach focuses on assessment of experimental features that are subsequently evaluated for bias and sensitivity. We do not currently know whether, to what extent, or in what kind of situations these differences would lead to substantive differences among the groups in the interpretation of the results of a study, or a set of studies. Our continued collaboration and coordination will allow us to address these questions.

Another difference is seen in the use of additional categories in the framework. For example, ORoC and EPA-IRIS each include an "analysis" domain and a "sensitivity" domain (Cooper et al., 2016–in this issue) in the evaluation of human observational and experimental animal studies, and other groups include a domain for "other" risk-of-bias issues, which is used to address topic-specific elements not otherwise incorporated in the tool. For example, in research questions that address multigenerational exposures, OHAT would add a separate question to assess whether or not there were appropriate methods to control for litter effects in experimental studies with developmental exposure as part of the "other potential threats to internal validity" question in its risk of tool. This analysis approach controls for the possibility that fetuses from a given litter might exhibit a similar response to a chemical exposure. The variety of issues that are included in "other" categories reflects the need for topic-specific considerations and the need to further develop, test, and refine the application of these risk-of-bias frameworks to environmental health assessments.

There is debate within the field of systematic review on whether and how to address potential funding or similar conflicts of interest by the study investigators. There has been some debate within the Cochrane Collaboration over the benefits and challenges of addressing funding source or conflict of interest within risk of bias as opposed to elsewhere in an evaluation (Bero, 2013; Dunn et al., 2014; IOM, 2009; Sterne, 2013). Some methods such as the Navigation Guide have a separate risk-of-bias question to assess conflict of interest. In support of this, the Navigation Guide approach has cited empirical data from various research fields that have demonstrated the ability of funding source to influence study outcome, from studies on health impacts of tobacco (Barnes and Bero, 1997, 1998), to safety and efficacy of pharmaceuticals (Bero et al., 2007; Lexchin et al., 2003; Lundh et al., 2012; Perlis et al., 2005), and medical procedures (Popelut et al., 2010). OHAT does not include a separate risk-of-bias question, but examines the potential influence of funding and conflict of interest as possible sources of heterogeneity and as part of evaluating potential publication bias across a body of evidence. GRADE considers the influence of funding under selective outcome reporting bias (within the risk of bias domain) or as a separate issue (i.e., publication bias) during assessment of the evidence (Guyatt et al., 2011b). ORoC and EPA-IRIS do not currently incorporate an evaluation of funding source or conflict of interest in their review process (but do evaluate the potential for selective reporting and publication bias).

## 3. Challenges and future directions

There are challenges in applying the systematic review approach originally designed for evaluation of randomized clinical trials to the more heterogeneous studies (observational epidemiology, experimental animal, and mechanistic studies) used to assess environmental exposures. These challenges are interrelated and include concerns about the need to: ensure common understanding of terminology and definitions; evaluate study limitations or strengths that span more than one domain; correctly characterize complex issues within a structured approach; develop approaches to address mechanistic data; and develop empirical data on the importance of individual risk-of-bias domains.

Table 4
Consideration of domains across five methods for assessing risk-of-bias of experimental animal studies.

| Domain | GRADE[a] | OHAT[b] | Navigation Guide | RoC | EPA-IRIS |
|---|---|---|---|---|---|
| Study design (including randomization, and allocation concealment prior to assignment, and experimental conditions) | • Considers randomization<br>• Considers lack of allocation concealment<br>• Considers if those enrolling subjects are aware of the group (or period in a crossover trial) to which the next enrolled subject will be allocated | • Assesses whether animals were assigned to treatment groups (including controls) using an explicit method to ensure randomization<br>• Evaluates whether personnel allocating animals to groups were unaware of the treatment groups until after animals were assigned treatments<br>• Considers if housing, husbandry, and treatment vehicle were identical across treatments | • Evaluates whether a random component was utilized to ensure the sequence of allocation to study group is unpredictable<br>• Evaluates whether allocation to groups was concealed from study investigators before and up until allocation assignment | • Corresponding domain is study design<br>• Considers randomization<br>• Allocation concealment not addressed<br>• Assesses control group and use of vehicle or sham treatment<br>• Considers age of animals when relevant | • Considers randomization<br>• Considers allocation concealment to treatment groups and endpoint evaluation groups<br>• Considers lack of control for other variables (e.g., surgery, animal husbandry) |
| Blinding during study | • Considers lack of blinding of caregivers and those administering exposure, conducting analysis, etc. | • Assesses whether research personnel were blind to treatment group during study | • Blinding during study and for outcome assessors assessed in a single question | • Not addressed | • Considered in the context of study design |
| Attrition/ exclusion | • Assesses incomplete accounting of subjects and outcome events<br>• Considers loss to follow-up and failure to adhere to the intention-to-treat principle in superiority trials | • Considers whether loss of animals was adequately addressed/documented when subjects removed from study or analysis | • Assesses incomplete outcome data—whether there is missing data, due to exclusion during the study or the analysis, that might introduce bias if reasons are related to the true outcome | • Considered under study design [statistical power for overall survival and loss of animals from studies and exposure conditions (for treatment related survival)] | • Considered in the context of analysis |
| Exposure | • Considers balanced exposure (intervention)<br>• Assesses full reporting of the exposure | • Evaluates whether purity and stability of treatment compound was assessed | • Assesses risk of exposure misclassification<br>• Accuracy, validity, and QA/QC of exposure assessment methods | • Considers chemical characterization, dose formulation, stability and delivery (some overlap with confounding)<br>• Dosing regimen (level, frequency, number of dose levels) and exposure duration (some overlap with sensitivity) | • Assesses test article composition, purity, stability, source<br>• Considers administration methods (vehicle and exposure condition controls; analytic protocol)<br>• Frequency and duration, not dose levels or spacing, considered under sensitivity |
| Outcome assessment | • Considers if those recording outcomes, those adjudicating outcomes, or data analysts are aware of the arm to which subjects are allocated<br>• Also addressed in selective outcome reporting | • Evaluates whether outcome was assessed with valid and reliable methods applied consistently (e.g., same method and time)<br>• Also assesses whether outcome assessors were blinded to treatment | • Assesses risk of outcome misclassification<br>• Assesses whether outcomes were assessed and defined consistently across all study participants, using valid and reliable measures<br>• Blinding during study and of outcome assessors assessed in a single question | • Assesses adequacy and consistency of pathology procedures (e.g., necropsy, gross pathology, histology, or diagnosis)<br>• Overlaps with sensitivity | • Assesses blinding of evaluators<br>• Considers sampling process (e.g., sufficient number of slides or trials)<br>• Assesses reliability/validity of protocols, including protocol controls |
| Selective reporting | • Assesses selective outcome reporting bias<br>• Considers incomplete or absent reporting of some outcomes and not others on the basis of the results | • Assesses whether results were provided for all relevant pre-specified measures and subjects | • Assesses whether outcome data is reported for all animals for all pre-specified primary and secondary outlines in the pre-specified manner | • Provide results for all relevant measures | • Assesses if results provided for all relevant measures and tested animals<br>• Important details reported (e.g., maternal health in pup analyses; lesion severity) |
| Conflict of interest | • Issues related to conflict of interest assessed on many levels: publication bias, selective outcome reporting bias, and through conflict of interest management at the evidence to decision level | • Addresses conflict of interest elsewhere in methods outside risk of bias | • Assesses whether financial conflict of interest is present for any of the study authors | • Not addressed | • Not addressed |
| Analysis | • Addressed under risk of bias when study level data is used but ideally dealt with through re-analysis of original data | • Analysis and statistical methods addressed under other potential threats to internal validity | • Analysis and statistical methods addressed under domain for other | • Considers appropriate combination of findings and of statistical models | • Accounts for early deaths or unexpected complications<br>• Considers decisions for results presentation (e.g., dichotomized or |

(continued on next page)

Table 4 (continued)

| Domain | GRADE[a] | OHAT[b] | Navigation Guide | RoC | EPA-IRIS |
|---|---|---|---|---|---|
| | | | | | continuous data) and analysis (e.g., statistical unit choice) |
| Sensitivity | • Not addressed | • Exposure and outcome sensitivity addressed under questions Exposure and Outcome assessment | • Exposure and outcome sensitivity addressed in Exposure and Outcome assessment domains | • Sensitivity questions are part of the questions in some of the other categories.<br>• Study design: suitability of the animal model and statistical power<br>• Exposure: exposure level and duration<br>• Outcome: observation duration (e.g., to allow for sufficient latency) | • Assesses suitability of animal model (e.g., strain) and group size<br>• Assesses suitability of exposure and endpoint evaluation timing (animal age; time of day), latency, frequency, and duration<br>• Considers endpoint evaluation sensitivity and specificity (e.g., positive and negative controls) |
| Other | • Considers other limitations:<br>◦ Stopping early for benefit<br>◦ Use of un-validated outcome measures<br>◦ Carryover effects in crossover trial | • Questions added on a project-specific basis to address potential threats to internal validity not addressed elsewhere (e.g., appropriate statistics, use of litter as unit of analysis) | • Considers:<br>◦ Early stopping due to data-dependent process<br>◦ Claim of fraudulence<br>◦ Atypical deviation from study methods<br>◦ Selective reporting of subgroups<br>◦ Insensitive instrument to measure outcomes | • Potential for confounding (e.g., contaminants, animal husbandry conditions)<br>• Route of exposure considered as a factor for external validity. | • Option to add questions or considerations specific to the situation being assessed<br>• Evaluations of the statistical methods, and the exposure levels and spacing tested, are considered in separate, subsequent steps |

[a] Assessment occurs on an outcome basis for each study and then, across studies for a specific question; See Guyatt et al. (2011c) for details on the GRADE approach to assessing internal validity.

[b] See http://ntp.niehs.nih.gov/go/38673 for the current OHAT risk-of-bias tool.

### 3.1. Communication across differences in study quality terminology

The focus of this paper is on the assessment of risk of bias of individual studies and how this assessment can be used (e.g., as a factor used to assess certainty in the body of evidence) in systematic reviews addressing environmental health questions. Risk of bias is one specific aspect of the larger concept of "study quality"; the definition of study quality can vary widely across the fields of systematic review and environmental health. There are a number of terms used preferentially by different groups to address concepts that can be addressed as part of study quality (e.g., see terminology discussion in Viswanathan et al., 2012). Therefore an important aspect of transparency in a systematic review is to clearly outline how study quality is assessed within an evaluation, where in the review process study quality is assessed, and how the assessment of individual studies, as well as the body of evidence, are considered in reaching final conclusions on the overall body of evidence. Inconsistent use of "study quality" creates the potential for miscommunication, so to improve clarity this paper presents and defines some of the common terms used in environmental health reviews and publications (see Table 1).

As seen by the variations among the five groups discussed here, the term "risk of bias", even when defined as above, can encompass different concepts or elements. For example, two of the groups include a domain relating to study "sensitivity" or the ability of the study to detect a true risk or hazard (Cooper et al., 2016–in this issue) for example, evidence of substantial exposure (e.g., level, duration, frequency, or probability) during an appropriate exposure window; a range of exposure levels or duration of exposure which allows for evaluation of exposure-response relationships; and an adequate length of follow-up in cohort studies; some of these features may also be incorporated into the other domains. Some, but not all, of the elements relating to study sensitivity can be included in other risk of bias domains (e.g., exposure); other attributes are not "biases" per se (e.g., exposure level or range encompassed by the study population), and so may require additional modification of terminology to facilitate communication across groups.

### 3.2. Risk-of-bias issues can be considered in more than one domain

One of the major challenges to harmonization of tools across groups is that several issues can reasonably be considered in more than one risk-of-bias domain or in different phases of the process for integrating evidence. For example, healthy worker hire effects and healthy worker survival effects are biases that can be considered as selection bias or potential confounder(s) in a risk-of-bias tool. Similarly, appropriate duration (or timing) of exposure for the effect in question that is reported in animal studies could be considered under sensitivity in the IRIS-EPA and ORoC approaches, indirectness in the GRADE framework, or exposure as a risk of bias element. In animal studies, treatment levels could be considered as either a risk of bias or sensitivity element. The evaluation of the potential for confounding, which is a separate domain in most risk-of-bias approaches, may overlap with several risk-of-bias domains (such as selection bias and analyses) and may also be considered in steps subsequent to the risk-of-bias evaluation (e.g., interpretation of the study's findings). In most cases there is no "correct" domain to consider the issue (and each assessment may handle them differently), the important point is to be transparent on how the issue will be considered and not to rate or address the issue in more than one domain (AHRQ, 2013; Viswanathan et al., 2012).

### 3.3. Evaluating complex issues within a structured approach

While an aim of systematic review methodology is to increase the transparency and reproducibility of environmental health assessments (Thayer et al., 2014), these approaches must always include scientifically sound judgments. This is not unique to environmental health; however, it may be more challenging with a diverse mixture of study

designs with consideration of complex issues. Well-constructed structured approaches that allow for flexibility and that capture the scientific issues that arise in the design, conduct, and analysis of environmental observational studies and animal toxicology studies can increase the transparency of the process for reaching hazard assessment conclusions without being too proscriptive or introducing a systematic bias. A study's potential for bias falls along a continuum of risk, not within discrete categorical ratings of each risk-of-bias domain (Balshem et al., 2011). In the evaluation of complex issues, such as exposure assessment, differences between studies may not be adequately represented by using a categorical rating system, and the process may be strengthened by a more thorough description of issues encountered in the evaluation.

### 3.4. Application to mechanistic data

The consideration of mechanistic data as part of the evidence base in a systematic review presents challenges for identifying relevant studies, evaluating risk of bias, and in developing confidence statements on the body of evidence (regarding either the level of mechanistic data support for observed human or animal health effects, or that the mechanistic data provides evidence for a health effect in the absence of human epidemiological or experimental animal studies). The cellular, biochemical, and molecular events (or mechanistic data) that are relevant to evaluating the health effect(s) target of the review should be identified in consultation with experts on the health effect and chemical/exposure in question. Ideally the scope of the mechanistic data is identified in the protocol; however, if new endpoints or potential mechanisms are identified in the course of a review, the search should be expanded to include additional intermediate endpoints and pathways. There are no published approaches for assessing risk of bias for mechanistic studies with an in vitro exposure regimen. General "study quality" tools, such as ToxRTool (ECVAM, 2009; Schneider et al., 2009), are available; however, such methods often provide a mixed assessment of reporting quality and study conduct and present results as a summary score for each study which does not account for relative difference in the importance of certain factors (i.e., the Klimisch score). There are also no broadly accepted frameworks for reaching confidence ratings for use of mechanistic data in decision making, and thus there is a need for research efforts to gain experience and develop methods in this area. The National Academy of Sciences noted the lack of guidance for assessing and using mechanistic data, and recommended that relevance of in vitro studies for hazard assessment should include consideration of the relevance of the cell system used, the exposure concentrations, metabolic capacity of the test system, and the relationship between the in vitro response and a clinically relevant outcome measure (NRC, 2014).

### 3.5. Development of empirical evidence

The goal of a risk-of-bias assessment is to assess potential sources of bias that could reduce the credibility or certainty of the study findings. There is empirical evidence to support some "domains" or separate aspects of risk of bias (e.g., a lack of randomization can bias results away from the null toward larger effect sizes for randomized controlled trials) (reviewed in Higgins and Green, 2011). Fewer sources of bias have been investigated with data from experimental animal studies, but there is growing evidence for the importance of some domains from these studies as well (e.g., lack of randomization also increased effect sizes in animal studies, see Hirst et al., 2014; reviewed in NRC, 2014). Other domains are included primarily from support based on toxicological or epidemiological principles. Unfortunately, there is a lack of empirical evidence to inform the relative importance of domains or even which domains should be included when evaluating the findings from studies with a particular design (Balshem et al., 2011; Viswanathan et al., 2012). Developing this evidence base would help to refine and target risk-of-bias evaluations to capture the elements that have the greatest impact on study credibility.

### 3.6. Future developments

Members of the five groups represented in this paper began meeting informally in the fall of 2014 because of common interests in understanding, developing, or refining methods for assessing the credibility of individual studies as part of reaching conclusions on specific environmental health questions. We developed this paper to present our current approaches to risk-of-bias assessment and to highlight commonalities as well as differences across these methods. However, it is also important to highlight ongoing efforts as more researchers apply systematic review methods to environmental health data. Current focus areas for future development are listed below:

- Many of the groups have released or are developing systematic review "handbooks" for their organizations that describe their methodology including how to assess risk of bias of individual studies and how these assessments are ultimately used to inform conclusions.
- All of the groups are applying their approaches to specific environmental health research questions or case studies, and expect to use the knowledge gained in considering future methods refinements.
- The application of GRADE to studies of environmental health has been identified as a research priority by the GRADE Working Group (Morgan et al. 2016–in this issue).
- Many of the groups are working to test and refine emerging software tools for data extraction, risk-of-bias assessment, and analysis.

Increasing adoption of rigorous and standardized tools to assess potential sources of bias adds to the transparency and objectivity in the critical appraisal of evidence used to develop conclusions in literature-based evaluations. There is a great deal of active communication and methods development on approaches to assess risk of bias across groups applying systematic review approaches to environmental health questions. Members of these five groups began discussions in an effort to foster understanding and harmonization as risk-of-bias methods continue to evolve. One goal as we move forward is to develop empirical evidence to support individual risk-of-bias questions (or their removal). Empirical evidence and experience will allow the groups to go beyond comparing methods and to begin to explore the potential implications of the differences in the risk of bias approaches described in this document. As we gain that experience, the groups can ask whether or not methodological differences in the risk of bias approach would result in meaningful differences in interpreting study results or reaching conclusions in a systematic review. And, if so, the groups can make more informed decisions on modifying their risk of bias practices and the potential advantages or consequences of harmonizing methods. Another goal is to maintain communication across groups facing similar challenges. In conducting their own specific reviews, each group expects to develop and refine their methods for assessing the credibility of the study results. Continued dialogue will promote harmonization as different organizations and researchers adapt their own methods, as applicable to their research goals.

## Competing financial interests

The authors declare they have no actual or potential competing financial interests.

## Acknowledgements

We appreciate the comments and input received during the development of this manuscript from colleagues in the different groups

represented in this paper. Review and advice provided by April Luke, Warren Casey, Dave Allen, Neepa Choksi, and John Bucher were particularly helpful in clarifying language in the final manuscript. The views expressed are those of the authors and do not necessarily reflect the policies of the US Environmental Protection Agency.

## References


Agerstrand, M., Kuster, A., Bachmann, J., Breitholtz, M., Ebert, I., Rechenberg, B., et al., 2011. Reporting and evaluation criteria as means towards a transparent use of ecotoxicity data for environmental risk assessment of pharmaceuticals. Environ. Pollut. 159, 2487–2492.

AHRQ (Agency for Healthcare Research and Quality), 2013. AHRQ training modules for the systematic reviews methods guide. Available: http://www.effectivehealthcare.ahrq.gov/index.cfm/tools-and-resources/slide-library/ (accessed 11 October 2013).

Atkins, D., Eccles, M., Flottorp, S., Guyatt, G.H., Henry, D., Hill, S., et al., 2004. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC Health Serv. Res. 4, 38.

Bal-Price, A., Coecke, S., 2011. Guidance on Good Cell Culture Practice (GCCP). In: Aschner, M., Suñol, C., Bal-Price, A. (Eds.), Cell Culture Techniques Neuromethods vol. 56. Humana Press, New York, NY, pp. 1–25.

Balshem, H., Helfand, M., Schunemann, H.J., Oxman, A.D., Kunz, R., Brozek, J., et al., 2011. GRADE guidelines: 3. Rating the quality of evidence. J. Clin. Epidemiol. 64, 401–406.

Barnes, D.E., Bero, L.A., 1997. Scientific quality of original research articles on environmental tobacco smoke. Tob. Control. 6, 19–26.

Barnes, D.E., Bero, L.A., 1998. Why review articles on the health effects of passive smoking reach different conclusions. J. Am. Med. Assoc. 279, 1566–1570.

Bero, L.A., 2013. Why the Cochrane risk of bias tool should include funding source as a standard item. Cochrane Database Syst. Rev. 12, ED000075.

Bero, L., Oostvogel, F., Bacchetti, P., Lee, K., 2007. Factors associated with findings of published trials of drug–drug comparisons: why some statins appear more efficacious than others. PLoS Med. 4, e184.

Beronius, A., Molander, L., Ruden, C., Hanberg, A., 2014. Facilitating the use of non-standard in vivo studies in health risk assessment of chemicals: a proposal to improve evaluation criteria and reporting. J. Appl. Toxicol. 34, 607–617.

Birnbaum, L.S., Thayer, K.A., Bucher, J.R., Wolfe, M.S., 2013. Implementing systematic review at the National Toxicology Program. Environ. Health Perspect. 121, A108–A109.

Blair, A., Stewart, P., Lubin, J.H., Forastiere, F., 2007. Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. Am. J. Ind. Med. 50, 199–207.

Christensen, K., Christensen, C.H., Wright, J.M., Galizia, A., Glenn, B.S., Scott, C.S., et al., 2014. The use of epidemiology in risk assessment: challenges and opportunities. Hum. Ecol. Risk. Assess. Int. J. 21, 1644–1663.

Cooper, G.S., Lunn, R.M., Ågerstrand, M., Glenn, B.S., Kraft, A.D., Luke, A.M., et al., 2016. Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures. Environ. Int. 92-93, 605–610 (in this issue).

Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., et al., 2012. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. J. Epidemiol. Community Health 66, 1182–1186.

Dunn, A.G., Arachi, D., Hudgins, J., Tsafnat, G., Coiera, E., Bourgeois, F.T., 2014. Financial conflicts of interest and conclusions about neuraminidase inhibitors for influenza: an analysis of systematic reviews. Ann. Intern. Med. 161, 513–518.

ECVAM (European Centre for the Validation of Alternative Methods), 2009e. ToxRTool — Toxicological data Reliability Assessment Tool. Available: http://ecvam.jrc.it/page_pdf.cfm?voce = s&idvoce = 110

EFSA (European Food Safety Authority), 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA J. 1–90 (Available: http://www.efsa.europa.eu/en/efsajournal/pub/1637.htm accessed 15 October 2013).

Ghio, A.J., Sobus, J.R., Pleil, J.D., Madden, M.C., 2012. Controlled human exposures to diesel exhaust. Swiss Med. Wkly. 142, w13597.

Guyatt, G.H., Oxman, A.D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., et al., 2011a. GRADE guidelines: 7. Rating the quality of evidence–inconsistency. J. Clin. Epidemiol. 64, 1294–1302.

Guyatt, G.H., Oxman, A.D., Montori, V., Vist, G., Kunz, R., Brozek, J., et al., 2011b. GRADE guidelines: 5. Rating the quality of evidence–publication bias. J. Clin. Epidemiol. 64, 1277–1282.

Guyatt, G.H., Oxman, A.D., Vist, G., Kunz, R., Brozek, J., Alonso-Coello, P., et al., 2011c. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias). J. Clin. Epidemiol. 64, 407–415.

Higgins, J., Green, S., 2011. Cochrane Handbook for Systematic Reviews of Interventions. Available: www.cochrane-handbook.org (accessed 3 February 2013).

Hill, A.B., 1965. The environment and disease: association or causation? Proc. R. Soc. Med. 58, 295–300.

Hirst, J.A., Howick, J., Aronson, J.K., Roberts, N., Perera, R., Koshiaris, C., et al., 2014. The need for randomization in animal trials: an overview of systematic reviews. PLoS One 9, e98856.

IARC (International Agency for Reserach on Cancer), 1990. Cancer: Causes, Occruances and Control. No. 100. IARC Scientific Publications, Lyon, France, p. 352.

IOM (Institute of Medicine), 2009. Conflict of Interest in Medical Research, Education, and Practice. The National Academies Press, Washington, DC, p. 414 (Available: http://www.nap.edu/download.php?record_id = 12598# accessed 23 May 2015).

IOM (Institute of Medicine), 2011. Finding What Works in Health Care: Standards for Systematic Reviews. The National Academies Press, Washington, DC, p. 318 (Available: http://www.nap.edu/openbook.php?record_id = 13059 accessed 3 May 2013).

Krauth, D., Anglemyer, A., Philipps, R., Bero, L., 2014. Nonindustry-sponsored preclinical studies on statins yield greater efficacy estimates than industry-sponsored studies: a meta-analysis. PLoS Biol. 12, e1001770.

Krauth, D., Woodruff, T., Bero, L., 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. Environ. Health Perspect. 121, 985–992.

Lexchin, J., Bero, L.A., Djulbegovic, B., Clark, O., 2003. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. Br. Med. J. 326, 1167–1170.

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gotzsche, P.C., Ioannidis, J.P., et al., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 6, e1000100.

Lundh, A., Sismondo, S., Lexchin, J., Busuioc, O.A., Bero, L., 2012. Industry sponsorship and research outcome. Cochrane Database Syst. Rev. 12, MR000033.

McPartland, J., Lam, J., Lanier-Christensen, C., 2014. A valuable contribution toward adopting systematic review in environmental health. Environ. Health Perspect. 122, A10.

Mollenhauer, M.A., Bradshaw, S.G., Fair, P.A., McGuinn, W.D., Peden-Adams, M.M., 2011. Effects of perfluorooctane sulfonate (PFOS) exposure on markers of inflammation in female B6C3F1 mice. J. Environ. Sci. Health, Part A: Tox. Hazard. Subst. Environ. Eng. 46, 97–108.

Morgan, R.L., Thayer, K.A., et al., 2016. GRADE: Assessing the quality of evidence in environmental and occupational health. Environ. Int. 92-93, 611–616 (in this issue).

NRC (National Research Council), 2014. Review of EPA's Integrated Risk Information System (IRIS) Process. Press TNA, Washington, DC, pp. 1–154 (Available: http://www.nap.edu/openbook.php?record_id = 18764 accessed 5 January 2015).

NTP (National Toxicology Program), 2013a. Webinar on OHAT approach for systematic review. (September 26, 2013. Available: http://ntp.niehs.nih.gov/go/404906 accessed 28 January 2014).

NTP (National Toxicology Program), 2013b. Informational meeting on the draft OHAT Approach for Systematic Review and Evidence Integration for Literature-based Health Assessments. (April 23, 2013. Available: http://ntp.niehs.nih.gov/go/38751 accessed 28 January 2014).

NTP (National Toxicology Program), 2015a. Handbook for Preparing Report on Carcinogens Monographs - July 2015. Carcinogens OotRo, RTP, NC (Available: http://ntp.niehs.nih.gov/go/rochandbook accessed 20 July 2015).

NTP (National Toxicology Program), 2015b. OHAT Risk of Bias Rating Tool for Human and Animal Studies - January 2015. Office of Health Assessment and Translation, RTP, NC (Available: http://ntp.niehs.nih.gov/go/38673 accessed 25 Jan 2015).

NTP (National Toxicology Program), 2015c. Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence — January 9 2015. Office of Health Assessment and Translation, RTP, NC (Available: http://ntp.niehs.nih.gov/go/38673 accessed 25 Jan 2015).

Perlis, R.H., Perlis, C.S., Wu, Y., Hwang, C., Joseph, M., Nierenberg, A.A., 2005. Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. Am. J. Psychiatry 162, 1957–1960.

Popelut, A., Valet, F., Fromentin, O., Thomas, A., Bouchard, P., 2010. Relationship between sponsorship and failure rate of dental implants: a systematic approach. PLoS One 5, e10274.

Rich, D.Q., Liu, K., Zhang, J., Thurston, S.W., Stevens, T.P., Pan, Y., et al., 2015. Differences in birth weight associated with the 2008 Beijing Olympic air pollution reduction: results from a natural experiment. Environ. Health Perspect. (advanced publication).

Rooney, A.A., Boyles, A.L., Wolfe, M.S., Bucher, J.R., Thayer, K.A., 2014. Systematic review and evidence integration for literature-based environmental health science assessments. Environ. Health Perspect. 122, 711–718.

Rothman, K.J., Greenland, S., 2005. Causation and causal inference in epidemiology. Am. J. Public Health 95 (Suppl. 1), S144–S150.

Schneider, K., Schwarz, M., Burkholder, I., Kopp-Schneider, A., Edler, L., Kinsner-Ovaskainen, A., et al., 2009. "ToxRTool", a new tool to assess the reliability of toxicological data. Toxicol. Lett. 189, 138–144.

Schünemann, H.J., Oxman, A.D., Higgins, J.P.T., Vist, G.E., Glasziou, P., Guyatt, G.H., et al., 2012. Chapter 11: presenting results and 'summary of findings' tables. In: JPT, H., Green, S. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions Version 510 [updated March 2011]. The Cochrane Collaboration.

Sterne, J.A., 2013. Why the Cochrane risk of bias tool should not include funding source as a standard item. Cochrane Database Syst. Rev. 12, ED000076.

Sterne, J., Higgins, J., Reeves, B., on behalf of the development group for ACROBAT-NRSI, 2014. A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI), Version 1.0.0. (Available: http://www.riskofbias.info accessed 28 September 2014).

Thayer, K.A., Wolfe, M.S., Rooney, A.A., Boyles, A.L., Bucher, J.R., Birnbaum, L.S., 2014. Intersection of systematic review methodology with the NIH reproducibility initiative. Environ. Health Perspect. 122, A176–A177.

Viswanathan, M., Ansari, M., Berkman, N.D., Chang, S., Hartling, L., McPheeters, L.M., et al., 2012. Assessing the Risk of Bias of Individual Studies When Comparing Medical Interventions. Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Publication No. 12-EHC047-EF. Available: http://www.effectivehealthcare.ahrq.gov/index.

cfm/search-for-guides-reviews-and-reports/?pageaction = displayproduct&productid = 998 accessed 3 January 2013).
WHO (International Programme on Chemical Safety, World Health Organization), 1999l. Principles for the Assessment of Risks to Human Health From Exposure to Chemicals. IPCS Environmental Health Criteria 210, Geneva, p. 312 (Available: http://www.inchem.org/documents/ehc/ehc/ehc210.htm).
Woodruff, T.J., Sutton, P., 2014. The Navigation Guide Systematic Review Methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environ. Health Perspect. 122, 1007–1014.